UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Modelo Aditivo Generalizado com função baseada num Perceptrão Multicamada: Estimação da função razão de possibilidades e respectivo intervalo de confiança

Teresa Martins Gonçalves

Mestrado em Bioestatística

Trabalho de Projeto orientado por:
Professora Doutora Lisete Sousa
Professor Doutor Carlos Geraldes

2021

Nota

Este trabalho é redigido segundo as regras do Acordo Ortográfico da língua portuguesa de 1990.

Agradecimentos

Quero começar por agradecer aos meus orientadores, Professora Doutora Lisete Sousa e Professor Doutor Carlos Geraldes, por todo o apoio e disponibilidade demonstrados. À minha família por me ter dado força, nesta fase tão difícil, principalmente à minha mãe, que sempre acreditou em mim, à minha mana por ter sido sempre uma inspiração para mim, ao meu mano que, apesar de mais novo, todos os dias me dizia que tinha que acabar este tormento e à minha avó, que mesmo longe está sempre comigo. Aos meus amigos, por estarem constantemente ao meu lado. Um especial agradecimento à Jebedo, que sempre foi a minha companheira de estudo, risadas e choro, durante todo o percurso do mestrado, à Tinha, que nunca me deixou desistir, à Caldeirinha, que me apoiou em todas as circunstâncias e à Guida, que fez noitadas comigo por videochamada, enquanto lia um livro.

Resumo

Este trabalho tem como objetivo analisar o desempenho de um Modelo Aditivo Generalizado (GAM) com função de ligação não paramétrica baseada num Perceptrão Multicamada (MLP) em comparação com um GAM com função de ligação logística. O racional na utilização deste tipo de função de ligação prende-se com o facto de um MLP ser um aproximador universal a uma função contínua. Assim, é expectável que utilizado como função de ligação de uma GAM, este seja suficientemente flexível para encontrar a solução mais adequada a um bom desempenho tendo em conta os dados que serviam para o treino do modelo. Adicionalmente, é desenvolvido um algoritmo que implementa uma função de razão de possibilidades (odds ratio) no âmbito de variáveis contínuas, bem como a obtenção dos respectivos intervalos de confiança. Esperava-se obter resultados importantes em estudos na área da saúde.

Os dados utilizados neste trabalho foram recolhidos na UCI de um hospital português e 3 dias após a admissão foi avaliado o desfecho do doente. Numa primeira fase foi feita uma revisão da literatura relativamente à temática das metodologias em perspectiva neste estudo, nomeadamente: Modelos Aditivos Generalizados (GAM), função de ligação não paramétrica, Perceptrão Multicamada, função de razão de possibilidades e estimação de intervalos de confiança. Recorrendo a um conjunto de dados clínicos, efectuou-se uma análise descritiva dos dados e, com estes, estimou-se um GAM com função de ligação logística e outro com função de ligação não paramétrica, baseada num Perceptrão Multicamada. Posteriormente, implementou-se um algoritmo para estimar a função de razão de possibilidades para uma variável explicativa contínua. Uma vez que, para variáveis binárias, o método de estimação era similar ao que era utilizado no caso de um Modelo Linear Generalizado (GLM). Simultaneamente foi feito um estudo comparativo entre dois métodos de obtenção de intervalos de confiança da função de razão de possibilidades implementada.

Palavras Chave: Modelo Linear Generalizado, Modelo Aditivo Generalizado, Razão de Chances, Perceptrão Multicamada.

Abstract

This work analyzes the performance of a Generalized Additive Model (GAM) with a non-parametric link function based on a Multilayer Perceptron (MLP) compared to a GAM with a logistic link function. The rationale for using this type of link function is that a MLP is an universal approximator to a continuous function. Thus, it is expected that when used as a link function for a GAM, it will be flexible enough to find the most suitable solution for good performance given the data used to train the model. Additionally, an algorithm should be developed to implement an odds ratio function for continuous variables, as well as to obtain the respective confidence intervals. This development is important in the context of epidemiological studies.

The data used in this study were collected in the ICU of a Portuguese hospital and the patient's outcome was assessed 3 days after admission. In a first stage, a literature review will be made regarding the methodologies in perspective in this study, namely: Generalized Additive Models (GAM) Non-parametric link function, Multi-layer Perceptron, odds ratio function and estimation of confidence intervals. Using a set of clinical data, a descriptive analysis of the data is performed and, with these, a GAM with a logistic link function and another with a non-parametric link function, based on a Multilayer Perceptron, is estimated. Later on, an algorithm will be implemented to estimate the odds ratio for continuous variable. Since, for binary variables, the estimation method is similar to the one used in the case of a Generalized Linear Model (GLM). A comparative study will be also made between two methods of obtaining confidence intervals for the odds ratio function.

Keywords: Generalized linear model, Generalized additive model, Odds Ratio, Multilayer perceptron.

Acrónimos e Siglas

AUC - Area Under The Curve

GAM - Generalized additive model

GLM - Generalized linear model

MLP - Perceptrão Multicamada

MSE - Mean squared error

OR - Odds ratio

ROC - Receiver operator characteristic

UCI - Unidade de cuidados intensivos

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O desenvolvimento tecnológico revolucionou a ciência, com a chegada de processos que optimizaram os métodos de investigação. A manipulação de conjuntos de dados de grandes dimensões tornou-se um processo possível e fiável, recorrendo à criação e aplicação de métodos estatísticos, o que tornou muito mais rápida e fácil a análise de dados (Cabral, 2019a). A recolha e a análise de dados é um processo fundamental para que se obtenham resultados úteis para o problema ao qual se pretende dar resposta, assim como a utilização de métodos estatísticos robustos.

Os métodos estatísticos são uma das ferramentas com maior importância nos avanços na área da medicina. As técnicas de sumariação de características amostrais são conhecidas como análise descritiva. A análise descritiva apresenta uma grande importância numa primeira fase de um estudo. Permite a obtenção de um quadro onde é possível extrair informações de medidas de tendência, dispersão e localização, bem como a representação gráfica da distribuição das componentes em estudo (Bazak *et al.*, 2013).

O estudo que se vai iniciar, nesta tese, é caracterizado por uma estrutura transversal, uma vez que todas as medições foram realizadas num único momento. Para que um estudo resulte é necessário que seja definida a questão à qual se pretende dar resposta, qual a população a estudar, qual o método de recolha e análise de dados a aplicar, as variáveis em estudo e os métodos a aplicar. Os estudos tranversais são utilizados quando se pretende descrever as características de uma determinada população no que se relaciona com variáveis que podem influenciar, ou não, o comportamento dos objectos de estudo e os seus padrões de distribuição. A definição da variável dependente, ou resposta, e das variáveis independentes, ou explicativas, é essencial no desenho de estudos com uma estrutura transversal. No presente caso, o conjunto de dados é formado por uma variável resposta binária e quatro variáveis explicativas quantitativas contínuas. Estes dados e estas variáveis foram recolhidos numa unidade de cuidados intensivos de um hospital português e dizem respeito à temperatura corporal, pressão arterial, frequência cardíaca e potássio. Neste estudo, a variável resposta representa a morte 3 dias após a entrada do

paciente na Unidade de Cuidados Intensivos de um hospital português.

A área deste trabalho, área da saúde, é uma das que mais recorre à estatística para encontrar respostas. A área da saúde está em contacto constante com a estatística. Dado ser uma área de grande responsibilidade, todas as decisões têm de ser corroboradas a partir de estudos rigorosos, por estar em causa a vida ou a sua qualidade. Para que as decisões sejam tomadas da melhor forma, são utilizadas métricas com base em modelos de regressão. Os modelos Linear Generalizado e Aditivo Generalizado são frequentemente escolhidos no suporte para a tomada de decisão.

Os Modelos Lineares Generalizados (GLM) podem ser comparados à estrutura específica de uma rede neuronal denominada por Perceptrão Simples. As redes neuronais possuem a característica de reconhecer padrões, ou outros tipos de relações, nos dados em estudo. No geral, uma rede neuronal, pode ser vista como sendo uma rede constituída por neurónios e sinapses artificiais. Um neurónio é uma função que soma todos os sinais de entrada e aplica uma função de ativação, sendo as mais usuais, a logística ou a tangente hiperbólica. Uma sinapse serve de interligação entre 2 neurónios e aplica um peso ao sinal que a atravessa.

Neste trabalho, o Modelo Aditivo Generalizado será aplicado com o objectivo de se encontrar o modelo mais parcimonioso por forma a dar resposta ao problema em questão, ou seja, perceber não só qual a probabilidade de morte dos doentes três dias após darem entrada na Unidade de Cuidados Intensivos (UCI), mas também quais os factores que têm uma maior influência na variável resposta, morte/sobrevivência do indivíduo após 3 dias nos UCI.

Em medicina, grande parte das decisões a tomar tem como objectivo recuperar a saúde do doente ou mesmo salvar-lhe a vida. Por norma, não são envolvidas muitas variáveis e o que se procura é extrair informação relevante para se perceber o estado de gravidade do doente, além de quantificar o risco de morte e/ou obter uma estimativa de mortalidade intra-hospitalar (Geraldes, 2017).

Este trabalho contempla a aplicação de um estudo de simulação que visa comparar o desempenho de um modelo aditivo generalizado com função de ligação logística e de um modelo aditivo generalizado com função de ligação flexível, baseado num Perceptrão Multicamada (MLP). Um MLP é um sistema semelhante a um perceptrão simples, mas pode ser constituído por uma ou mais camadas intermédias e ter mais que um neurónio de saída.(Chaphalkar *et al.*, 2015)

Posteriormente, será efectuada a estimação da função de razão de possibilidades a partir de cada uma das abordagens do modelo aditivo generalizado (GAM), referidas anteriormente. Esta análise será feita com base na adaptação de um algoritmo utilizado para a estimação da função de razão de possibilidades para as variáveis explicativas: temperatura corporal, pressão sanguínea, frequência cardíaca e potássio sérico.

Para finalizar, e dado que se pretende a aplicação e comparação de duas abordagens para o modelo aditivo generalizado, utilizam-se as medidas de adequabilidade AUC e Erro Quadrático Médio (MSE), para perceber qual das abordagens resultou melhor em cada uma das variáveis explicativas e para os dois modelos seleccionados.

Para a concretização do estudo recorreu-se ao *software R* com o código correspondente em Apêndice.

1.1 Objectivos

Neste trabalho, pretende-se a aplicação dos GAM com duas funções de ligação a uma amostra recolhida numa unidade de cuidados intensivos (UCI), respeitante a 519 doentes.

A análise de dados na área de Medicina é essencial para que se tomem decisões acertadas e para que seja possível extrair o máximo de informação útil a partir dos dados (Geraldes, 2017).

Esta dissertação tem como objectivo o estudo do desempenho do Modelo Aditivo Generalizado (GAM) com função de ligação logística, em comparação com um GAM com função de ligação não paramétrica baseada num Perceptrão Multicamada (MLP).

A utilização de uma função de ligação não paramétrica baseada num MLP apresenta a vantagem do MLP ser um aproximador universal a uma função contínua. Desta forma, é expectável que, utilizado como função de ligação de um GAM, este seja suficientemente flexível para encontrar a solução mais adequada para os dados de treino do modelo.

Adicionalmente, deverá ser implementado um algoritmo que utiliza uma função de razão de possibilidades (*odds ratio*), no âmbito de variáveis contínuas, bem como a obtenção dos respectivos intervalos de confiança. Este desenvolvimento é importante no contexto dos estudos da área da saúde.

Capítulo 2

Metodologia

Ao longo deste capítulo será referida e explicada a metodologia aplicada ao longo do projecto.

2.1 Modelo de Regressão Linear Múltipla

Os modelos de regressão linear são caracterizados pela presença de uma variável resposta Y, e uma, ou mais, variáveis explicativas, $X_1, X_2, ..., X_p$. Estes modelos são muito utilizados em estatística e assumem a forma

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon \tag{2.1}$$

Na expressão 2.1, Y representa variável resposta, $\beta_0, \beta_1, ..., \beta_p$ são parâmetros desconhecidos, $X_1, ..., X_p$ são as variáveis explicativas e ε é a componente aleatória. (Cabral, 2019b).

No entanto, nem sempre o modelo de regressão linear se adequa visto que é complicado encontrar uma relação de linearidade entre a variável dependente, Y, e a variáveis independentes, $X_1, ..., X_p$. Desta forma, é sugerida a utilização de outros modelos que não tenham como requisito a linearidade entre a variável resposta e as variáveis explicativas. Existem modelos alternativos como, por exemplo, os GAM que não requerem que a relação entre os preditores e a variável resposta seja linear.

2.2 Modelo Linear Generalizado

O Modelo Linear Generalizado, também conhecido por GLM, surge como uma extensão do Modelo de Regressão Linear. Este modelo é caracterizado pela presença de uma variável aleatória Y conhecida como variável resposta ou variável dependente, e um vector $\mathbf{X} = (X_1, ..., X_p)^T$ de p variáveis explicativas também conhecidas como covariáveis ou variáveis explicativas. O objectivo é perceber de que forma as covariáveis conseguem explicar alguma da variabilidade de Y.

Os dados assumem a forma

$$(y_i, x_{i1}, \ldots, x_{ip}), i = 1, \ldots, n, \tag{2.2}$$

este resultado baseia-se na concretização de $(Y, \mathbf{X})$ numa amostra com n indivíduos ou unidades experimentais. O vector $\mathbf{Y}$ apresenta componentes independentes Y_i, sendo o vector $\mathbf{Y} = (Y_1, ..., Y_n)^T$.

Os dados podem ser apresentados na seguinte forma matricial:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} x_{11} & x_{12} & ... & x_{1p} \\ x_{21} & x_{22} & ... & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & ... & x_{np} \end{pmatrix} \tag{2.3}$$

Numa amostra, é frequente encontrarmos muitos índividuos que contêm o mesmo vector de covariáveis, sobretudo com variáveis explicativas de natureza qualitativa. Consequentemente a matriz X apresenta grupos de linhas idênticas e, por conseguinte, os dados podem ser agrupados, (Turkman *et al.*, 2000).

Os índividuos podem ser agrupados em g grupos diferentes, por forma a que os n_j índividuos do grupo j partilhem o mesmo vector de covariáveis, com $j = 1, ..., g$, $g < n$, $\sum_{j=1}^{g} n_j = n$ e $\mathbf{x^*}_j = (x_{j1}, ..., x_{jp})^T$. A representação matricial dos dados é então:

$$\overline{\mathbf{y}} = \begin{pmatrix} \overline{y_1} \\ \overline{y_2} \\ \vdots \\ \overline{y_g} \end{pmatrix} \quad X_g^* = \begin{pmatrix} x_{11}^* & x_{12}^* & ... & x_{1p}^* \\ x_{21}^* & x_{22}^* & ... & x_{2p}^* \\ \vdots & \vdots & & \vdots \\ x_{g1}^* & x_{g2}^* & ... & x_{gp}^* \end{pmatrix}, \tag{2.4}$$

com $\overline{y_j}$, j = 1,...,g, a representar a média da variável resposta dos indivíduos que pertencem ao j-ésimo grupo, sem linhas idênticas em X_g^*.

É de salientar que o agrupamento de dados tem especial interesse quando se trabalha com covariáveis somente de natureza qualitativa.

No caso dos GLM a relação entre a combinação linear das variáveis explicativas e a váriável resposta Y, possa contemplar distribuições que não só a Normal. Deste modo, é considerada qualquer distribuição desde que pertença à família exponencial de distribuições (Geraldes, 2017).

As distribuições pertencentes à família exponencial apresentam, para dados não agrupados, a função densidade ou massa de probabilidade definida :

$$f(y|\theta, \phi) = e^{\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}}, \tag{2.5}$$

onde: θ é um parâmetro escalar e representa a forma canónica do parâmetro de localização; ϕ um parâmetro escalar que representa a dispersão e usualmente conhecido. As funções $a(.), b(.), c(.)$, são funções reais conhecidas e $b(.)$ diferenciável.

Um GLM costrói-se a partir de duas componentes: a componente aleatória e a componente sistemática ou estrutural. Na componente aleatória, Y é considerada uma

Tabela 2.1: Distribuições, respectiva função de ligação canónica, domínio e variância de Y

Distribuição de Y	Função de ligação canónica	Domínio de Y	Var(Y)
Normal $N(\mu, \sigma^2)$	identidade (μ)	$]-\infty, +\infty[$	σ^2
Poisson $(Pois(\lambda))$	logarítmica$(\log(\lambda))$	$\{0, 1, ...\}$	λ
Gama $(Ga(\nu, \frac{\nu}{\mu}c))$	recíproca $(-\frac{1}{\mu})$	$(0, +\infty)$	$\frac{\mu^2}{\nu^2}$

variável com distribuição pertencente à família exponencial. Desta forma, dado o vetor de covariáveis $\mathbf{x}_i$, as variáveis Y_i são condicionalmente independentes e pertencem à família exponencial, com $E(Y_i|\mathbf{x}_i) = \mu_i$, $i = 1, ..., n$ e com um parâmetro de dispersão que não depende de i.

A componente sistemática tem como base o preditor linear definido por:

$$\eta = X\beta \tag{2.6}$$

onde

$$\mathrm{X} = \begin{pmatrix} 1 & x_{11} & ... & x_{1p} \\ 1 & x_{21} & ... & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & ... & x_{np} \end{pmatrix}, \tag{2.7}$$

X define-se como uma matriz de especificação e $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$ é um vector de parâmetros de dimensão p+1.

O valor médio $E[Y_i|\mathbf{x_i}] = \mu_i$ relaciona-se com o preditor linear através de uma função de ligação $h(.)$ monótona e diferenciável. Para o *i-ésimo* indivíduo o valor esperado traduz-se em:

$$\mu_i = h(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}) = h(\eta_i) \tag{2.8}$$

de forma inversa obtém-se

$$g(\mu_i) = \eta_i \tag{2.9}$$

onde $g(.) = h^{-1}(.)$ representa a inversa da função de ligação, a qual se designará igualmente por função de ligação.

Quando existe coincidência entre o preditor linear e o parâmetro canónico, ou seja, quando $\eta_i = \theta_i$ com $\theta_i[1, ..., x_{i1}, ..., x_{ip}]^T\beta)$, a função de ligação é designada canónica.

Na tabela (2.1), são referidas algumas distribuições pertencentes à família exponencial com as respectivas funções de ligação canónicas.

Contudo, quando $Y \in]0, +\infty[$, nem sempre a função canónica é a melhor opção. Deste modo, deve ser utilizada uma função de ligação cujos valores esperados sejam sempre positivos.

A título de exemplo são apresentadas duas funções de ligação. Uma para um GLM com variável reposta contínua e outra para um GLM com variável resposta discreta.

- **GLM com variável resposta contínua**

Modelo Normal

Na presença de uma variável resposta contínua é possível aplicar o Modelo de Regressão Normal.

Como apresentado na Tabela 2.1, para o modelo com distribuição de Y Normal a função de ligação é a função identidade que corresponde ao Modelo de Regressão Linear. Desta forma, tem-se:

$$g(\mu_i) = \mu_i\mu_i = \eta_i \tag{2.10}$$

com

$$\mu_i = \eta_i. \tag{2.11}$$

- **GLM com variável resposta discreta**

Modelo de Poisson

Na presença de uma variável resposta discreta, no caso de respostas em forma de contagens, é possível aplicar o modelo de Regressão de Poisson.

Como apresentado na Tabela 2.1, para o modelo Poisson a função de ligação é a função logarítmica.

Desta forma, tem-se:

$$g(\mu_i) = \log \mu_i \tag{2.12}$$

com

$$\mu_i = h(\eta_i) = e^{\eta_i}. \tag{2.13}$$

2.3 Modelo Aditivo Generalizado

O Modelo Aditivo Generalizado (GAM), entre outros, é muito importante para a tomada de decisão em Medicina. Permitem-nos estudar e interpretar o risco de mortalidade (Bulhosa, 2019).

Uma das vantagens que os GAM apresentam, comparativamente aos GLM, é o facto não ser exigida uma relação de linearidade entre a variável resposta e as variáveis explicativas.

A estimação de um GAM é inspirada no método de estimação de um GLM. Os GAM são definidos da seguinte forma:

$$\begin{gathered} E(Y|x_1, x_2, ..., x_p) = h(\beta_0 + f_1(x_1) + ... + f_p(x_p)) \\ com \\ g(\mu_i) = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) \end{gathered} \tag{2.14}$$

Na expressão (2.12), h(.) representa a função de ligação do modelo, $X_1, ..., X_p$ definem-se como variáveis explicativas e $f_j(.), j = 1, ..., p$, são as funções parciais, também conhecidas como funções suavizadoras.

Um dos exemplos de funções suavizadoras são os *splines*, $s(.)$, e são definidos por

$$s(x) = \sum_{k=1}^{p} b_k(x)\beta_k, \tag{2.15}$$

onde $b_k(x)$ é chamada função base de ordem k e β_k é o parâmetro desconhecido de ordem k. A função suavizadora pode, também, ser definida da seguinte forma

$$s(x) = \beta_0 + \sum_{k=1}^{p} x^k\beta_k \tag{2.16}$$

Na expressão (2.14), definiu-se a função suavizadora da expressão (2.13) a partir de uma regressão polinomial genérica de ordem p, com p a representar as variáveis explicativas.

Contudo, nem sempre a regressão polinomial genérica é a mais adequada, quando existem irregularidades na nuvem de pontos. Desta forma, recorre-se a uma regressão por *splines* definida por intervalos, com base em m pontos designados por "nós", $(c_1, ..., c_m)$. (Geraldes, 2017)

Para um ponto c_k genérico, recorre-se a uma função base definida por funções lineares "truncadas", restringidas a um intervalo, do tipo $(x - c_k)+$

$$(x - c_k)+ = \begin{cases} x - c_k, & x \geq c_k \\ 0, & x < c_k. \end{cases} \tag{2.17}$$

Através da combinação linear das funções "truncadas"presentes na expressão (2.15) obtém-se o *spline s(x)*

$$s(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{m} \beta_{1_k}(x - c_k) + . \tag{2.18}$$

Nem sempre é possível garantir a continuidade das funções base lineares. Desta forma, são consideradas funções não lineares em cada um dos intervalos e, finalmente, o *spline s(x)* assume a forma:

$$s(x) = \beta_0 + \beta_1 x + ... + \beta_p x^p + \sum_{k=1}^{m} \beta_{p_k}(x - c_k)^p + . \tag{2.19}$$

As funções base lineares não garantem que haja continuidade da primeira derivada da combinação linear dessas funções e, consequentemente, são consideradas funções não lineares em cada um dos intervalos. Deste modo, uma generalização da função suavizadora de grau p traduz-se em:

$$s(x) = \beta_0 + \beta_1 x + ... + \beta_p x^p + \sum_{k=1}^{m} \beta_{pk}(x - c_k)^p + . \tag{2.20}$$

Um dos métodos utilizados para se estudar a função *spline* é o método dos mínimos quadrados que é efectuado recorrendo à minimização de:

$$\|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|^2, \tag{2.21}$$

com $\beta = (\beta_0, \beta_1, ..., \beta_p, \beta_{p1}, ..., \beta_{pm})^T$.

Um dos procedimentos, para tal, consta em controlar o grau de suavização do *spline* a partir de um termo penalizador que se adiciona a $\|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|^2$. A estimação de s(.) passa então pela minimização de:

$$\|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|^2 + \lambda \int [s''(t)]^2 dt. \tag{2.22}$$

O parâmetro $\lambda \in [0, +\infty[$ funciona como um mediador entre a adaptação do *spline* e o seu grau de suavização. Quando λ tende para zero, obtém-se um *spline* tal que $s''(x)$ apresente valores elevados e que permita construir um novo conjunto de dados a partir de outros previamente conhecidos. Se λ tender para $+\infty$ a função s coincide com a recta dos mínimos quadrados.

O parâmetro λ, também conhecido como parâmetro suavizador, pode ser estimado de forma optimizada a partir de métodos como a validação cruzada, por exemplo. Neste contexto, o parâmetro é estimado pela minimização de:

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{s}_\lambda^{-i}(\mathbf{x}_i))^2, \lambda > 0. \tag{2.23}$$

Na expressão supra, $\hat{s}_\lambda^{-i}$ é a estimativa da função *spline* em $\mathbf{x_i}$, ou seja, estimada a partir de todas as observações excepto a *i-ésima*. Calcula-se o valor da função de validação cruzada, *CV*, para os vários valores de λ e escolhe-se aquele que minimiza o valor da função (Lawrence *et al.*, 1998).

2.4 Estimação da Função de Razão de Possibilidades para uma Variável Explicativa Contínua

Ao longo deste trabalho será aplicado o algoritmo de razão de possibilidades (OR) a um GAM, não paramétrico com função de ligação logística. Quando se está perante um estudo que pretende verificar a associação entre a exposição e um efeito é frequentemente utilizado o OR. O OR pode explicar a relação entre a exposição a um factor e o desenvolvimento de uma doença, quando se sabe *a priori* que o efeito corresponde a doença.

Neste contexto, o OR representa a possibilidade da ocorrência de doença, dada a exposição a um determinado factor X, comparativamente à possibilidade da ocorrência da mesma doença, dada a ausência de exposição ao mesmo factor X (Cadarso *et al.*, 2005).

Desta forma, $\pi_{d,x} = P(Y = 1|X = x)$ exprime a probabalidade de ocorrência da doença.

$$\Omega_{dx} = \frac{\pi_{dx}}{1 - \pi_{dx}}. \tag{2.24}$$

A expressão (2.23), é utilizada no exemplo do delineamento de um estudo caso-controlo. Seleccionam-se dois grupos de indivíduos, o grupo dos casos e o grupo dos controlos. O grupo dos casos é constituído pelos elementos que contraíram a doença e o grupo dos controlos é formado pelos elementos que não sofrem da doença em estudo. Deste modo, $\pi_{d,e} = P(Y = 1|X = 1)$ representa a probabilidade de doença no grupo dos expostos e $\pi_{d,ne} = P(Y = 1|X = 0)$ a probabilidade de doença no grupo dos não expostos.

Reunidas as condições anteriores, o OR pode ser definido por:

$$OR = \frac{\Omega_{d,e}}{\Omega_{d,ne}} = \frac{\pi_{d,e}/(1 - \pi_{d,e})}{\pi_{d,ne}/(1 - \pi_{d,ne})}, \tag{2.25}$$

onde $\Omega_{d,e}$ traduz a possibilidade de doença nos indivíduos expostos e $\Omega_{d,ne}$ a possibilidade de doença nos indivíduos não expostos.

Para se estimar a função OR (expressão (2.26)) para uma ou mais variáveis explicativas contínuas e para um GAM com função de ligação flexível não paramétrica foi implementado um algoritmo onde o OR aplicado à k-ésima covariável X_k num ponto genérico x, em comparação com um ponto $x_0 \in [min(x_k), max(x_k)]$, de acordo com a figura 2.1:

$$OR_k^{x_0}(x) = E_{X_k}\left[\frac{\pi(X_1, ..., x, ..., X_p)/(1 - \pi(X_1, ..., x, ..., X_p))}{\pi(X_1, ..., x_0, ..., X_p)/(1 - \pi(X_1, ..., x_0, ..., X_p))}\right] \tag{2.26}$$

Para mais detalhes consultar Cadarso *et al.*(2005).

2.4.1 Intervalos de Confiança para a Função de Razão de Possibilidades (função OR)

A estimação dos intervalos de confiança para a função OR (bandas de confiança) pode ser obtida através do método de Bootstrap. Este método de reamostragem pode ser aplicado quando se desconhece a distribuição de amostragem de um estimador, neste caso, desconhece-se a distribuição de amostragem do estimador de $OR_k^{x_0}(x)$ para cada valor de x.

Neste trabalho, a estimação dos intervalos de confiança para a função OR (bandas de confiança), para cada uma das variáveis preditoras, começa pela obtenção de B amostras

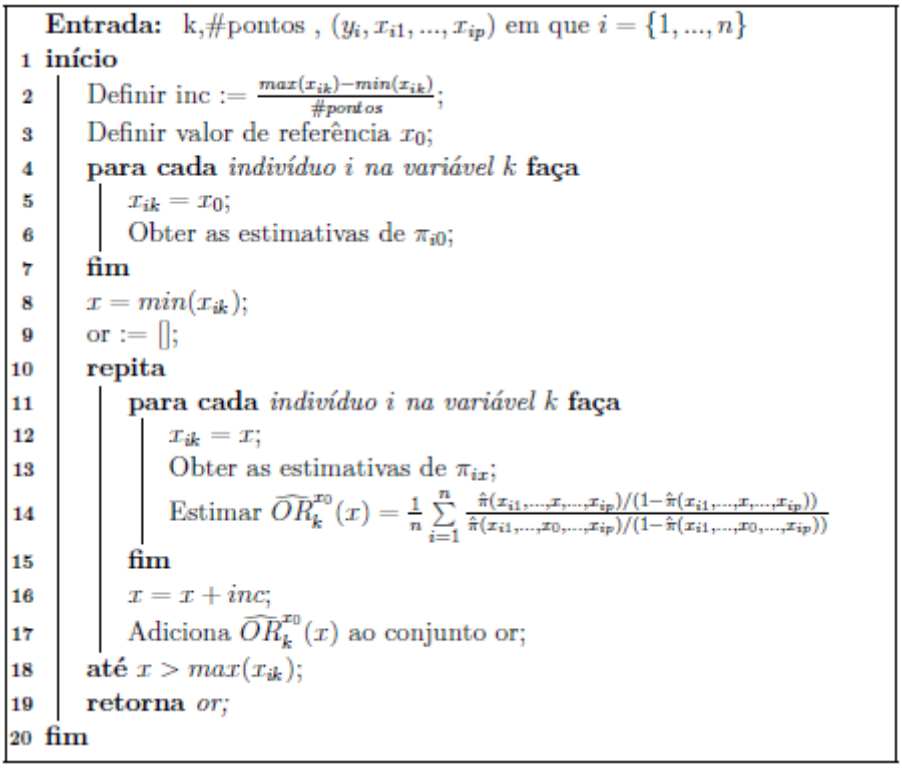

Figura 2.1: Algoritmo para a estimação da função de razão de possibilidades de uma variável explicativa contínua X_k (retirado de Geraldes (2017))

Bootstrap a partir da amostra orignal (com reposição). De seguida, treinam-se B modelos aditivos generalizados (com função de ligação MLP) originando B funções OR estimadas de acordo com o algoritmo da Figura 2.1.

O processo fica concuído após o cálculo dos percentis 25, 50 e 75 dos B valores de OR obtidos para cada valor x da variável preditora. Estes valores, da variável preditora, são definidos de acordo com uma grelha de pontos no intervalo de valores assumidos pela variável conforme se vê no algoritmo.

A representação gráfica dos 3 percentis nos pontos considerados da variável preditora, permite desenhar a função mediana do OR e as referidas bandas de confiança.

2.5 Perceptrão Multicamada

O Perceptrão Multicamada (MLP) é uma rede neuronal que se organiza em camadas de neurónios onde um neurónio de determinada camada está ligado a todos os neurónios da camada anterior. A primeira camada, denominada camada de entrada, tem tantos neurónios quanto o número de variáveis explicativas em estudo. As camadas intermédias, ou escondidas, contêm neurónios que apresentam uma sinapse extra que se liga a um neurónio com valor de saída igual a 1 designado por viés (Papoila e Rocha, 2011).

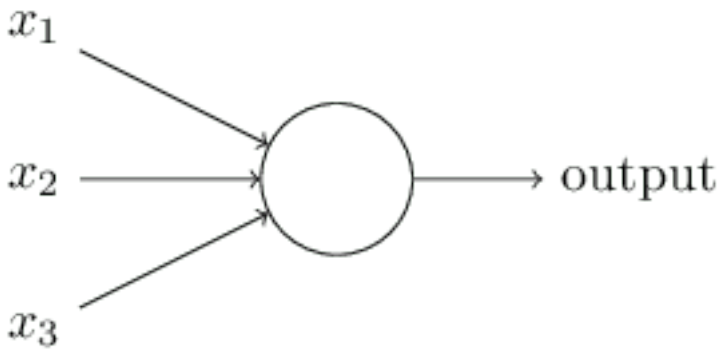

Figura 2.2: Ilustração de um perceptrão simples

Considere-se um MLP com uma camada escondida com q neurónios, uma camada de entrada com p neurónioss e uma camada de saída com um neurónio.

Um MLP com estas características apresenta uma maior flexibilidade aquando da sua aplicação a um modelo comparativamente ao Perceptrão Simples, Figura 2.2; mais detalhes em Geraldes (2017). Para o *i-ésimo* indivíduo considera-se:

$$\mu_i = h\left(\omega_0 + \sum_{j=1}^{q} \omega_j h\left(\omega_{0j} + \sum_{k=1}^{p} \omega_{jk} x_{ik}\right)\right), \tag{2.27}$$

com h a representar a função de ligação, ω_0, ω_{jk} e x_{ik} represetam constantes reais. Desta forma é possível modelar funções contínuas com apenas uma camada escondida.

Ao longo deste trabalho será aplicado o MLP com apenas uma camada escondida (Figura 2.3), sendo esta camada suficiente para o bom desempenho do processo, como se comprova pelo *Teorema da Aproximação Universal*, que mostra que é possível modelar qualquer função contínua com apenas uma camada escondida (Hornik *et al.*, 1989).

Terorema: Seja $\varphi(.)$ uma função contínua não constante, limitada e monótona crescente. Seja I_p um hipercubo unitário de dimensão p, $[0,1]^p$. $C(I_p)$ é denominado por espaço das funções contínuas em I_p. Desta forma, para qualquer função $f(.) \in C(I_p)$ e $\epsilon > 0$, existe um inteiro M e conjuntos de constantes reais α_k, γ_k, e ω_{jk} em que $j = 1, ..., q$ e $k = 1, ..., p$ tal que a função

$$F(x_1, ..., x_p) = \nu_0 + \sum_{k=1}^{p} \alpha_k \varphi\left(\sum_{j=1}^{q} \omega_{jk} x_j + \nu_k\right) \tag{2.28}$$

pode considerar-se como uma aproximação da função $f(.)$ onde

$$\left|F(x_1, ..., x_p) - f(x_1, ..., x_p)\right| < \epsilon, \tag{2.29}$$

para $x_1, x_2, ..., x_p \in I_p$.

Desta forma, tem-se uma camada de entrada com as covariáveis em estudo e um neurónio com valor um, uma camada escondida e uma camada de saída com a função μ e o neurónio viés (Bishop, 1996). Como referido, tem-se um MLP com p neurónios na camada de entrada, q neurónios na camada escondida e um neurónio na camada de saída. Considera-se que os pesos sinápticos de entrada de cada neurónio escondido se representam por $\omega_k = (\omega_{1k}, ..., \omega_{jk})$ e que os pesos do neurónio viés correspondem a $\nu = (\nu_0, ..., \nu_q)$. Os pesos sinápticos que interligam a camada escondida e a camada de saída são representados por $\alpha = (\alpha_1, ..., \alpha_q)$ (Lawrence, 1998).

A expressão (2.25) comprova que um MLP com apenas uma camada escondida é suficiente para a aproximação ϵ de uma função contínua, considerando o conjunto de covariáveis $x_1, ..., x_p$ e uma função objectivo $f(x_1, ..., x_p)$ (Geraldes, 2017).

2.6 Medidas de Adequabilidade do Ajuste do Modelo

2.6.1 AUC e Curva ROC

Para melhor aferir a adequabilidade do ajuste do modelo consideraram-se medidas como a área abaixo da curva ROC (AUC) e o Erro Quadrático Médio (MSE).

A curva ROC (*Receiver Operating Characteristic*) e a AUC são muito utilizadas para medir o desempenho de modelos de classificação. A curva ROC é uma forma de representação da relação entre a sensibilidade e a especificidade de um teste diagnóstico, cujos eixos tomam valores no intervalo $[0, 1]$ e é caracterizada pelo cálculo da proporção de verdadeiros positivos, designada sensibilidade, e de falsos positivos, designada especificidade. O cálculo das proporções é definido por:

$$\textit{Sensibilidade} = \frac{\text{Positivos classificados correctamente}}{\text{Total de positivos}} \tag{2.30}$$

$$\textit{Especificidade} = \frac{\text{Negativos classificados incorrectamente}}{\text{Total de negativos}}. \tag{2.31}$$

Estas proporções são utilizadas no cálculo da AUC, que é a probabilidade $P(W > Z)$, onde W representa a variável aleatória correspondente aos resultados para os casos e Z a corresponde aos controlos (Cortes *et al.*, 2003). É de notar que a AUC varia entre 0.5 e 1 e quanto mais próximo de 1 for o valor, melhor o modelo distingue entre os casos e os controlos. Consequentemente, é possível assumir que a AUC representa uma medida de classificação.

Teorema: Seja c um classificador fixo. Considere-se $w_1, .., w_m$ o resultado de c nos casos e $z_1, ..., z_n$ os resultados nos controlos. A AUC, A, associada a c é dada por:

$$A = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n'} 1_{w_i > z_j}}{mn'}, \tag{2.32}$$

que é o valor da estatística do teste de Mann-Whitney, w_i é um vector de observações dos casos, z_j é um vector de de observações dos controlos e m e n' as dimensões correspondentes a w_i e z_j, respectivamente, com $n' \leq m$.

2.6.2 Erro Quadrático Médio

O erro quadrático médio, MSE, é uma medida de qualidade do ajuste do modelo que que pode pode ser estimado a partir da expressão (2.32) nos casos de resposta binária:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.33}$$

com $(y_i - \hat{y}_i)^2$ a representar os resíduos ao quadrado. Esta medida coincide com o score de Brier,que é outra medida de calibração (Geraldes, 2017).

Quanto menor for o valor de MSE melhor será a adequabilidade do modelo. No entanto, quanto maior o número de parâmetros a estimar, maior a complexidade do modelo.

2.6.3 Teste de DeLong

O teste de DeLong é uma abordagem não paramétrica à análise de duas ou mais áreas sob curvas ROC correlacionadas. A área abaixo da curva ROC empírica calculada pela regra trapezoidal é igual à estatística de teste de Wilcoxon (teste não-paramétrico para comparação de duas amostras independentes). O resultado é uma matriz de covariância estimada. Considerando uma amostra de dimensão n, dos quais m sofrem o evento de interesse, denominado C_2, e n' indivíduos que não sofreram qualquer ocorrência do evento de interesse, denominado C_1. Considerando as definições de sensibilidade e especificidade, a probabilidade, θ, de seleccionar aleatoriamente uma observação da população representada por Z ser inferior ou igual a seleccionar aleatoriamente uma observação da população representada por W, é apresentada abaixo, como uma média sobre um kernel, ψ:

$$\hat{\theta} = \frac{1}{mn'} \sum_{j=1}^{n'} \sum_{i=1}^{m} \psi(W_i, Z_j), \tag{2.34}$$

com

$$\psi(W, Z) = \begin{cases} 1 & Z < W \\ \frac{1}{2} & Z = W. \\ 0 & Z > W \end{cases} \tag{2.35}$$

Generalizado a equação anterior, para K classificadores binários, C_1, W_i^k denota a probabilidade estimada de que pertence à classe 1. Z_j^k pode efinido para observações em C_2. Consequentemente, k-ésima AUC é calculada por:

$$\hat{\theta}^k = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \psi(W_i^k, Z_j^k). \tag{2.36}$$

$\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_K)^T \in \mathbb{R}^K$ representa as K AUC empíricas, $\theta = (\theta_1, \ldots, \theta_K)$ é o vector de AUCs verdadeiros e $\mathbf{L}$ é um vector de coeficientes. A hipótese nula assume a expressão: $H_0 : \theta_1 = \theta_2, \quad i.e. \quad \mathbf{L}^T\boldsymbol{\theta} = 0$

e a estatística de teste é dada por:

$$\frac{\mathbf{L}\hat{\boldsymbol{\theta}}^T - \mathbf{L}\boldsymbol{\theta}^T}{\left[\mathbf{L}\left(\frac{1}{m}\mathbf{S}_{10} + \frac{1}{n}\boldsymbol{S}_{01}\right)\mathbf{L}^T\right]^{\frac{1}{2}}} \sim \mathcal{N}(0, 1) \quad \text{sob a validade de} \quad H_0. \tag{2.37}$$

De notar que quando o valor-p associado ao teste de DeLong toma valores inferiores ao nível de significância considerado a hipótese H_0 é rejeitada. (Demler *et al.* 2012)

Capítulo 3

Resultados

Ao longo deste capítulo serão apresentados os resultados obtidos a partir dos métodos utilizados. Os dados foram recolhidos na UCI de um hospital português e 3 dias após a admissão foi avaliado o desfecho do doente. Inicialmente é feita uma análise descritiva e são apresentadas medidas de dispersão para cada variável, bem como gráficos que revelam o comportamento das variáveis explicativas. Os GLM, GAM logístico e GAM com função de ligação baseada num Perceptrão Multicamada são implementados para a modelação de uma variável resposta binária

Para o estudo que contou com 519 pacientes, foram consideradas uma variável resposta binária, morte, e quatro variáveis explicativas: temperatura corporal, pressão arterial, frequência cardíaca, potássio sérico.

3.1 Análise Exploratória dos Dados

3.1.1 Estatística Descritiva

A Tabela 3.1 apresenta o sumário das medidas de localização para cada uma das quatro variáveis explicativas, quando se verifica a morte na UCI, passados 3 dias de internamento, enquanto que a Tabela 3.2 sumaria informação análoga quando não há morte.

Na Tabela 3.2, apresentam-se as mesmas características amostrais, para o grupo de indivíduos que se encontravam vivos ao fim de 3 dias de internamento.

Tabela 3.1: Estatística descritiva relativa às variáveis explicativas à data da entrada na UCI, nos 196 pacientes que vieram a falecer ao fim de 3 dias.

	Mínimo	Quantil 0.25	Quantil 0.5	Média	Quantil 0.75	Máximo
Temp. corporal (ºC)	32.40	36	37.50	37.24	38.23	42
Press. arterial (mmHg)	30	54.25	84	90.80	104	227
Freq. cardíaca (bpm)	0	95.25	121.50	107.86	142	208
Potássio (mmol/l)	2	3.10	3.60	3.96	4.77	8.10

Tabela 3.2: Estatística descritiva relativa às variáveis explicativas à data da entrada na UCI, nos 323 pacientes que estavam vivos ao fim de 3 dias.

	Mínimo	Quantil 0.25	Quantil 0.5	Média	Quantil 0.75	Máximo
Temp. corporal (ºC)	33	36	37.30	37.06	37.80	40.30
Press. arterial (mmHg)	30	91.50	110	130.80	176.50	268
Freq. cardíaca (bpm)	11	91	117	110.20	132	220
Potássio (mmol/l)	1	3.20	3.60	3.83	4.30	9

Nos pacientes que vieram a falecer nos 3 dias seguintes, a temperatura média era ligeiramente superior, ao passo que a pressão arterial e frequência cardíaca médias são inferiores, sobretudo a pressão arterial. Os níveis médios de potássio são semelhantes.

- Verifica-se que a temperatura corporal para o grupo dos utentes que morrem após 3 dias internados na UCI varia entre os 32.40ºC e os 42ºC, sendo a média amostral de 37.24ºC; para o grupo dos utentes que sobrevivem após 3 dias internados na UCI varia entre os 33ºC e os 40.30ºC, sendo a média amostral de 37.06ºC.

- A pressão arterial para o grupo dos utentes que faleceram 3 dias após entrada na UCI apresenta valores que variam entre os 30mmHg e os 227mmHg. A média é de 90.80mmHg; para o grupo dos utentes que sobrevivem após 3 dias internados na UCI apresenta valores que variam entre 30mmHg e 268mmHg, sendo a média amostral 130.80mmHg.

- O ritmo cardíaco para o grupo dos utentes que morreram 3 dias após 3 dias na UCI mínimo é de 0 batimentos por minuto, o que significa possivelmente os indivíduos nestas condições faleceram aquando da entrada na UCI. Eventualmente, deveriam ter sido retirados da base de dados. O máximo é de 227 batimentos cardíacos por minuto. De acordo com a amostra, pode ainda concluir-se que, em média, o número de batimentos cardíacos por minuto é de, aproximadamente 107.86bpm; para o grupo dos utentes que sobreviveram 3 dias após entrada na UCI o mínimo é de 11bpm, o máximo de 220bpm e a média amostral é de 110.20bpm.

- Finalmente, para a variável potássio sérico para o grupo dos utentes que morreu 3 dias após entrada na UCI o valor mínimo é de 2mmol/l, o máximo é de 8.10mmol/l e a média amostral é de 3.96 mmol/l; para o grupo de utentes que sobreviveu 3 dias após a entrada na UCI o mínimo é de 1mmol/l, o máximo é de 9mmol/l e a média amostral de 3.83mmol/l.

3.1.2 Análise Gráfica

Optou-se pela representação gráfica a partir de histogramas e diagramas em caixa de bigodes, *boxplots*.

As distribuições de cada variável foram separadas em dois grupos: o grupo dos utentes que acabaram por falecer ("morte"=1) e o grupo dos que se mantinham vivos ao fim de 3 dias após entrada na UCI ("morte"=0).

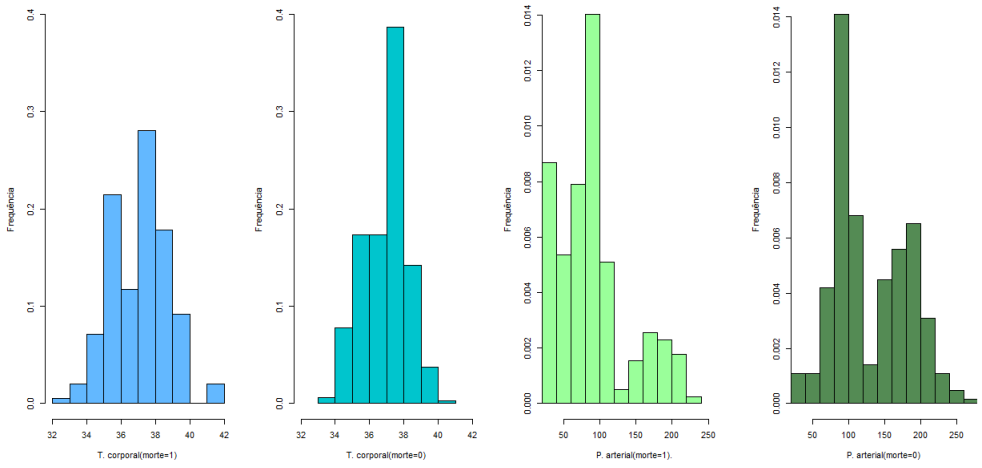

Figura 3.1: Histogramas das distribuições da temperatura corporal e pressão sanguínea tendo em conta o estado vital do paciente na UCI. A frequência no eixo yy representa a frequência relativa a dividir pela amplitude de classe.

Ao observar os histogramas das Figuras 3.1 e 3.2, parecem existir diferenças na distribuição e na dimensão dos grupos de indivíduos que sobreviveram e de indivíduos que morreram 3 dias após entrada na UCI, principalmente para a variável pressão arterial.

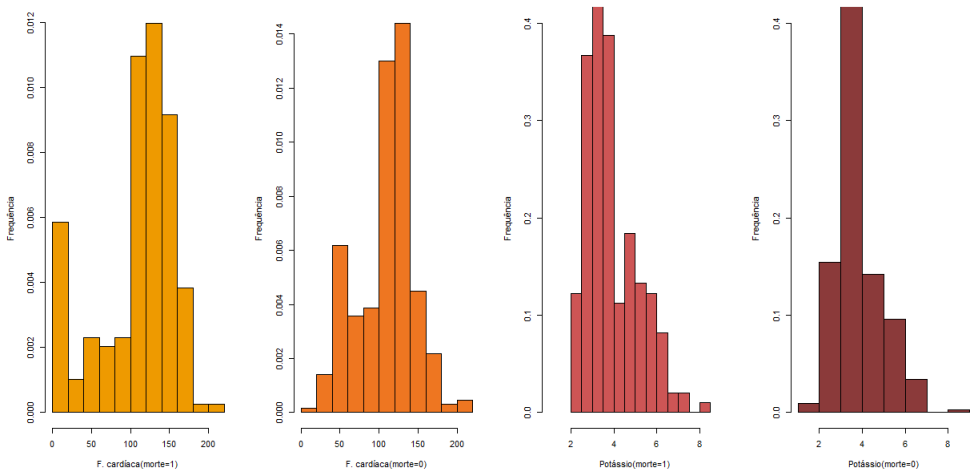

Figura 3.2: Histogramas das distribuições da frequência cardíaca e potássio séricotendo em conta o estado vital do paciente na UCI. A frequência no eixo yy representa a frequência relativa a dividir pela amplitude de classe.

Outra das representações também muito utilizadas para as variáveis contínuas é o diagrama caixa de bigodes. Infra, Figura 3.3, são representadas cada uma das variáveis explicativas em diagrama caixa de bigodes. De notar que a variável resposta é do tipo binário e quando asume o valor 0 representa vida e 1 representa morte.

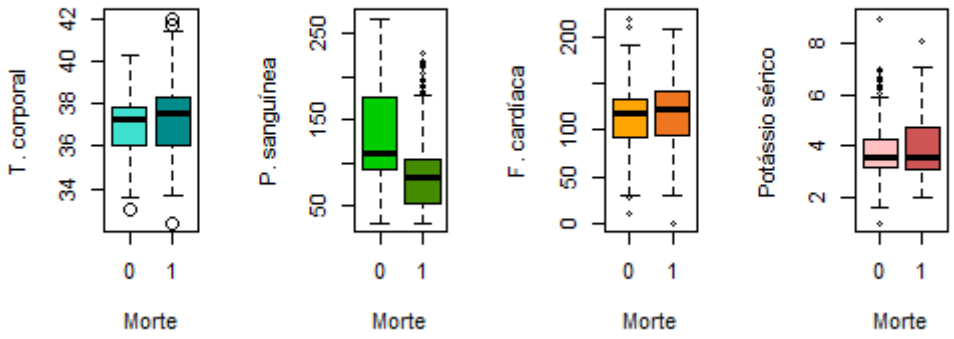

Figura 3.3: Diagramas em caixa de bigodes, para cada uma das distribuições associadas às variáveis explicativas (à esquerda de cada par apresenta-se o grupo dos sobreviventes e à direita o grupo dos que vieram a falecer).

A representação em diagramas em caixa de bigodes foi feita em função da variável resposta binária morte. Analisando os diagramas da pressão sanguínea, verdes, verifica-se

que as medianas são diferentes assim como a dispersão nos dois grupos, sendo ambas assimétricas à direita. No grupo onde se verifica morte os valores da pressão sanguínea são mais baixos do que no grupo dos sobreviventes. Nas restantes variáveis as medianas são próximas, embora também se verifiquem algumas diferenças ao nível da dispersão. Quanto à simetria das distribuições, nas restantes variáveis, mantém-se a tendência para distribuições assimétricas, com maior evidência na variável Frequência Cardíaca.

É possível constatar que em todos os diagramas em caixa de bigodes se apresentam *outliers*, sendo mais evidente para a variável no grupo "mortos"e para a variável Potássio Sérico no grupo "vivos".

Uma observação discrepante é definida como uma observação que apresenta um valor muito diferente dos valores das outras observações da amostra e que suscita dúvidas quanto ao que se passou com aquele elemento (Hanusz, 2016). Nos diagramas em caixa de bigodes, Figura 3.3., verifica-se, de forma mais evidente, que as amostras correspondentes às variáveis pressão arterial, e potássio sérico, são as que apresentam mais observações candidatas a *outlier*.

A correlação amostral é uma medida de associação linear entre cada par de variáveis em estudo (Cabral, 2019c). De salientar que a correlação de uma variável com ela mesma apresenta o valor um. Os valores do coeficiente de correlação variam no intervalo $[-1, 1]$ e neste caso recorreu-se ao coeficiente de correlação de *Pearson*. Por exemplo, se o coeficiente respeitante à associação linear entre duas variáveis for próximo de -1, então diz-se que as variáveis estão inversamente correlacionadas e que a correlação é forte, ou seja quanto maior o valor de uma variável menor será o da outra. Se o valor de associação entre as variáveis for próximo de zero, então as variáveis não estão correlacionadas. Se o valor do coeficiente assumir um valor próximo de um, a associação das variáveis é linear positiva, ou seja a correlação é forte, o que significa que, quanto maiores forem os valores de uma variável, maiores serão os da outra também. Em termos absolutos, considera-se que 0.9, ou mais, representa uma correlação muito forte; entre 0.7 a 0.9 considera-se uma correlação forte; 0.5 a 0.7 indica uma correlação moderada; 0.3 a 0.5 positivo ou negativo indica uma correlação fraca; e entre 0 e 0.3 indica uma correlação muito fraca, ou a ausência de correlação. (Mukaka, 2012)

Apresenta-se, posteriormente, a Tabela 3.3 com a matriz de correlações entre cada par de variáveis. Ao observar a Tabela 3.3, verifica-se que as variáveis que apresentam uma associação linear mais elevada são a temperatura corporal e a frequência cardíaca, (r=0.25). Contudo, o valor é muito reduzido, traduzindo-se numa correlação fraca.

Apesar de muito importante, a análise gráfica não é suficiente para se decidir que variáveis devem ser incluídas no modelo final. O objectivo é encontrar o modelo mais

Tabela 3.3: Matriz de correlações de Pearson para as variáveis explicativas

	Temp. corporal	Press. arterial	Freq. cardíaca	Potássio
Temp.corporal	1.00	-0.04	0.25	-0.03
Press. arterial	-0.04	1.00	0.08	-0.12
Freq. cardíaca	0.25	0.08	1.00	-0.04
Potássio sérico	-0.03	-0.12	-0.04	1.00

parcimonioso, por forma a explicar da melhor maneira a variável resposta.

3.2 Selecção do Modelo

Para se perceber quais as variáveis a considerar para o modelo final, foram construídos quatro modelos de regressão logística univariados para cada variável explicativo em estudo: temperatura corporal, pressão arterial, frequência cardíaca e potássio sérico. Para cada um dos modelos, apresenta-se um resumo dos resultados obtidos: o valor-p para os testes de hipóteses de Wald, cujas hipóteses são $H_0 : \beta_i = 0$ vs. $H_1 : \beta_i \neq 0$, (i=1,2,3,4).

Para a selecção das variáveis do modelo, foi considerado um nível de significância de 0.25, (Bishop 1996), sendo a hipótese nula rejeitada quando o valor-p é inferior ao nível de significância.

Começando pela variável temperatura corporal, o modelo de regressão logística univariado apresenta a forma:

$$g(\mu_1) = \beta_0 + \beta_1 \times (Temp.corporal), \tag{3.1}$$

e obtiveram-se os resultados apresentados na Tabela 3.4:

Tabela 3.4: Sumário do modelo de regressão logística univariado para a variável temperatura corporal

Coeficientes	Estimativa	Desvio-padrão	Est. Teste	Valor-p
(Constante)	-3.019	2.443	-1.65	0.1
Temperatura corporal	0.095	0.066	1.44	0.149

Observando a Tabela 3.4, verifica-se que a variável temperatura corporal é estatisticamente significativa para o modelo, uma vez que o valor-p associado ao teste é inferior ao nível de significância de 25%.

Para a variável pressão sanguínea, o modelo de regressão logística univariado apresenta a forma:

$$g(\mu_2) = \beta_0 + \beta_2 \times (Press.sanguinea), \tag{3.2}$$

cujo sumário é apresentado na Tabela 3.5,

Tabela 3.5: Sumário do modelo de regressão logística univariado para a variável pressão sanguínea

Coeficientes	Estimativa	Desvio-padrão	Est. Teste	Valor-p
(Constante)	1.348	0.244	5.532	<0.00001
Pressão sanguínea	-0.017	0.002	-7.776	<0.00001

Observando os resultados da Tabela 3.5, nomeadamente o valor-p, conclui-se que a variável pressão sanguínea é estatisticamente significativa para o modelo, uma vez que o valor-p é inferior ao nível de significância considerado para esta selecção, $\alpha = 0.25$.

Para a variável frequência cardíaca, o modelo de regressão logística univariado apresenta a forma:

$$g(\mu_3) = \beta_0 + \beta_3 \times (Freq.cardiaca), \tag{3.3}$$

e obtiveram-se os resultados apresentados na Tabela 3.6:

Tabela 3.6: Sumário do modelo de regressão logística univariado para a variável frequência cardíaca

Coeficientes	Estimativa	Desvio-padrão	Est. Teste	Valor-p
(Constante)	-0.355	0.25	-1.419	0.156
Frequência cardíaca	-0.001	0.002	-0.617	0.537

Observando os resultados na Tabela 3.6, nomeadamente o valor-p, conclui-se que a variável frequência cardíaca não é estatisticamente significativa para o modelo, uma vez

que o valor-p é superior ao nível de significância definido, $\alpha = 0.25$. No entanto, optou-se por se considerar esta variável no modelo final, uma vez que esta é estatisticamente significativa para o modelo na presença das outras 3 variáveis explicativas,

Para a variável potássio sérico, o modelo de regressão logística univariado apresenta a forma:

$$g(\mu_4) = \beta_0 + \beta_4 \times (Potassio), \tag{3.4}$$

e obtiveram-se os resultados apresentados na Tabela 3.7:

Tabela 3.7: Sumário do modelo de regressão logística univariado para a variável potássio sérico

Coeficientes	Estimativa	Desvio-padrão	Est. Teste	Valor-p
(Constante)	-0.978	0.341	-2.752	0.006
Potássio sérico	0.113	0.084	1.339	0.181

Observando os resultados da Tabela 3.7, nomeadamente o valor-p, conclui-se que a variável potássio sérico é estatisticamente significa para o modelo, uma vez que o valor-p é inferior ao nível de significância definido, $\alpha = 0.25$.

Para o modelo multivriável optou-se por considerar todas as variáveis explicativas. Verifica-se no entanto, que em alguns casos a forma funcional entre a variável explicativa e a variável resposta apresentam efeitos não lineares como é o caso da Frequência Cardíaca pelo que se deverá utilizar um GAM como modelo final. A expressão deste modelo é a descrita na equação 3.5 e será este o utilizado ao longo deste trabalho.

$$g(\mu) = \beta_0 + f_1(Temp.corporal) + f_2(Press.sanguinea) + f_3(Freq.cardiaca) + f_4(Potassio), \tag{3.5}$$

em que $g(.)$ representa uma função de ligação, $f_1, ..., f_4$ representam as funções suavizadoras univariadas, suavizadores de Kernel, também conhecidas como funções parciais e β_0 representa a ordenada na origem.

3.3 Modelo GAM logístico e GAM-MLP

Uma vez seleccionadas as variáveis para o modelo, serão aplicadas duas funções de ligação distintas ao GAM. A primeira será uma função de ligação logística e a segunda será uma função de ligação flexível baseada num MLP.

O objectivo é completar o modelo GAM com MLP com os valores obtido no GAM logístico e perceber se existem melhorias significativas.

Para perceber qual dos dois métodos seria mais adequado, entre o GAM com função logística e o GAM com função de ligação baseada num perceptrão multicamada (GAM-MLP) e calculou-se a AUC e o erro quadrático médio.

Como é possível observar na Tabela 3.8, as melhorias que se obtiveram no GAM com função logística para o modelo com função de ligação baseada num perceptrão multicamada não são muito significativas.

Tabela 3.8: Comparação dos valores das medidas de adequabilidade de ajuste para os modelos GAM logístico e GAM-MLP

Modelo	AUC	MSE
GAM função logística	0.782	0.178
GAM função de ligação MLP	0.787	0.175

Os dados da Tabela 3.8 foram obtidos com recurso ao $software R$, através do código desenvolvido para este trabalho e das respectivas bibliotecas (ver anexo). Concluiu-se que o número ideal de nós da camada escondida da rede neuronal é 34 tendo em conta a minimização do erro quadrático médio. Verificou-se que com esta configuração conseguimos obter uma AUC superior à que foi medida com o GAM logístico. No entanto, é possível constatar, que as melhorias não são estatisticamente significativas, após a realização do teste de DeLong, cujo valor-p obtido é 0.104. Não se rejeita a hipótese de igualdade de desempenho entre os dois modelos, aos níveis usuais de significância, e desta forma é possível concluir que o modelo GAM-MLP não traz melhorias significativas para a discriminação das categorias da variável resposta, quando comparado com o GAM logístico.

Posteriormente, são apresentados os resultados do estudo e análise das funções de razão de possibilidades.

3.3.1 Funções de razão de possibilidades e respectivos intervalos de confiança

Ao longo desta secção são apresentadas as funções OR para as variáveis pressão sanguínea e frequência cardíaca, considerando a mediana como valor de referência e foram obtidos os respectivos intervalos de confiança *Bootstrap* a 95%, considerando 50 amostras Bootstrap. Estas variáveis são as que se consideram clinicamente mais relevantes e que também

apresentam diferenças, ou seja, só para estas variáveis é que existe uma variação da função OR.

Temperatura Corporal

A função OR, para a temperatura corporal, foi obtida considerando a mediana da amostra da tensão arterial como valor de referência de 37.13^{o}C.

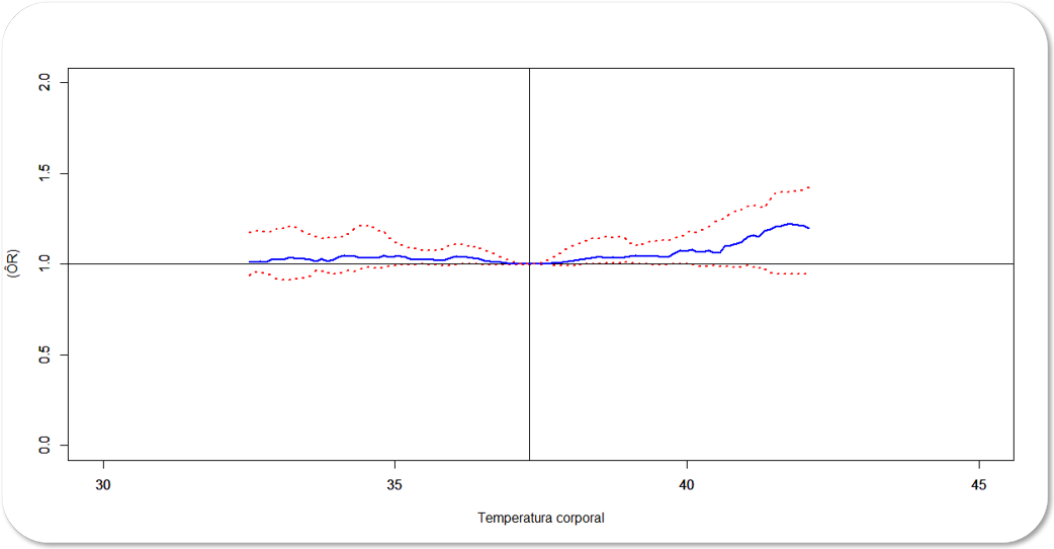

Figura 3.4: Estimativa da função parcial de razão de possibilidades e do respectivo intervalo de confiança a 95% para a temperatura corporal e a variável resposta morte

Ao observar a Figura 3.4, conclui-se que a possibilidade de morte, nos três dias seguintes à entrada para a UCI, aumenta com o aumento da temperatura corporal. Na amostra estudada, verificou-se que houve maior dificuldade em estabilizar os pacientes que chegaram à UCI com temperaturas muito elevadas.

Pressão Sanguínea

A função OR, para a pressão arterial, foi obtida considerando a mediana da amostra da tensão arterial, 99mmHg, como valor de referência.

Ao observar a Figura 3.5, conclui-se que a possibilidade de morte, três dias após a entrada para a UCI, aumenta com a diminuição da pressão sanguínea. A pressão sanguínea controla-se mais facilmente quando está elevada. Quer isto dizer que, é mais fácil fazer descer a pressão sanguínea, do que a fazer subir.

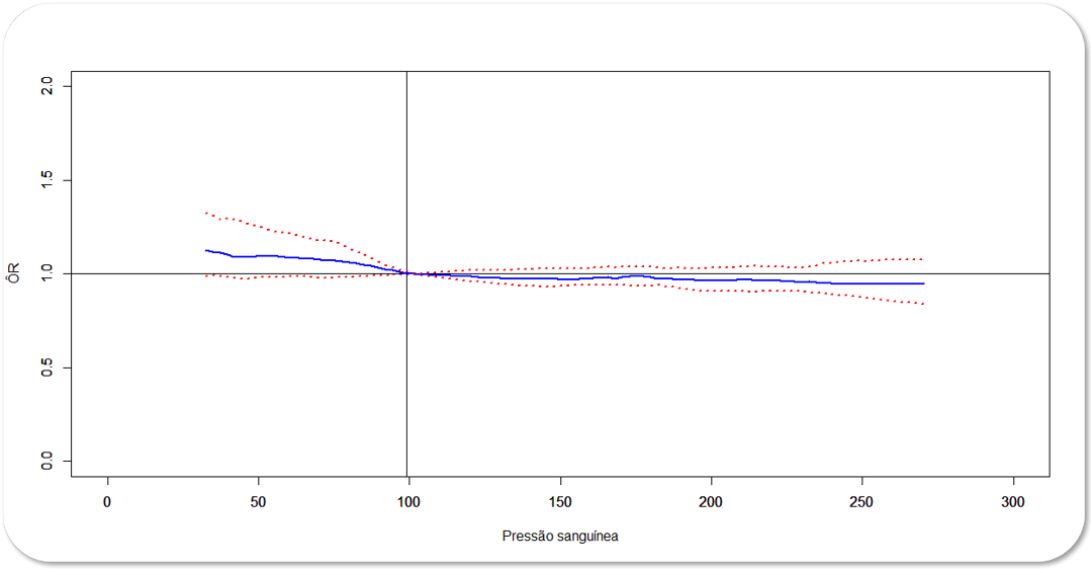

Figura 3.5: Estimativa da função parcial de razão de possibilidades e do respectivo intervalo de confiança a 95% para a pressão sanguínea e a variável resposta morte

Frequência Cardíaca

A função OR, para a frequência cardíaca, foi obtida considerando a mediana da amostra de batimentos cardíacos como valor de referência, 116 batimentos por minuto.

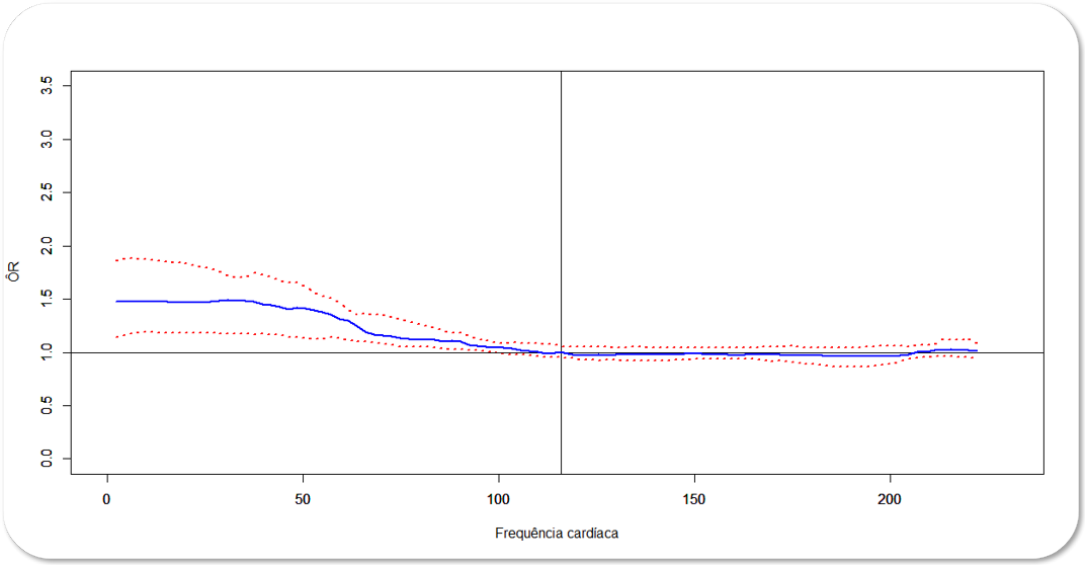

Figura 3.6: Estimativa da função parcial de razão de possibilidades e do respectivo intervalo de confiança a 95% para a frequência cardíaca e a variável resposta morte

De acordo com a Figura 3.6, pode concluir-se que a possibilidade de morte, três dias após a entrada para a unidade de cuidados intensivos, aumenta com a diminuição da frequência cardíaca. Tal como para a pressão sanguínea, verifica-se que é mais difícil estabilizar os pacientes quando estes apresentam um número de batimentos cardíacos muito baixo.

Potássio Sérico

A função OR, para a potássio sérico, foi obtida considerando a mediana da amostra, 3.87mmol/l, como valor de referência.

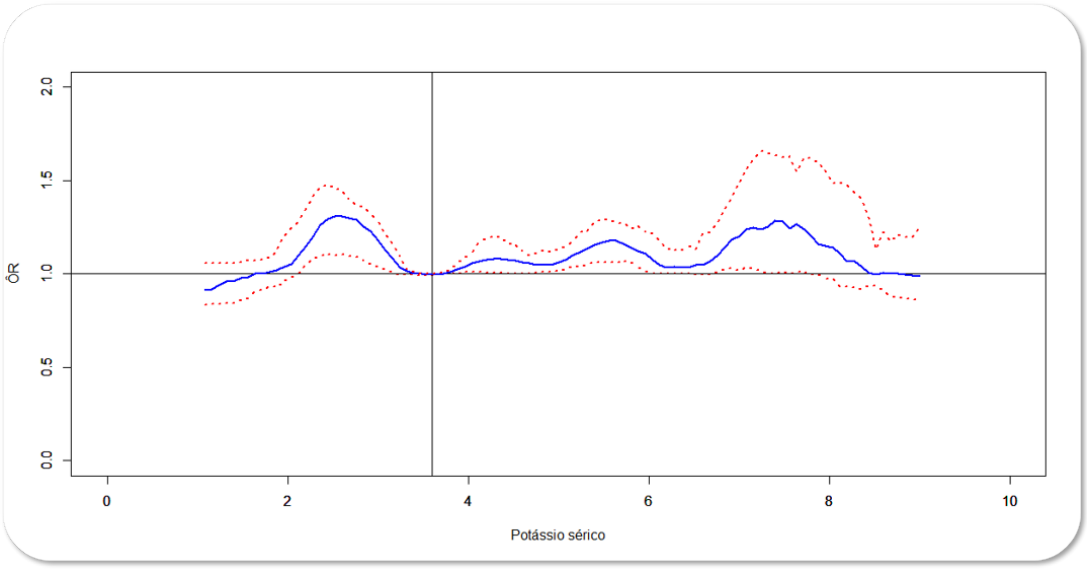

Figura 3.7: Estimativa da função parcial de razão de possibilidades e do respectivo intervalo de confiança a 95% para o potássio sérico e a variável resposta morte

De acordo com a Figura 3.7, pode concluir-se que a possibilidade de morte, durante os três dias seguintes à entrada na UCI, aumenta quando a variável toma valores baixos ou acima da mediana amostral. Existe uma tendência para o risco de morte aumentar quando os níveis de potássio aumentam ou diminuem. O facto do risco de morte voltar a aproximar-se de 1 para valores próximos do mínimo e do máximo da amostra, pode dever-se ao baixo número de pacientes com valores com esta ordem de grandeza.

Capítulo 4

Discussão e Conclusões

Ao longo deste trabalho foram comparados os desempenhos de modelos aditivos generalizados, um com uma função de ligação logística e outro com uma função de ligação baseada num MLP. Contudo, os objectivos não foram alcançados, pois o esperado era obter um desempenho significativamente melhor para o GAM-MLP.

Começou por ser feita uma análise individual de cada variável, para se perceber quais os valores críticos para um indivíduo correr risco de vida, tendo por base uma amostra de 519 indivíduos que deram entrada numa unidade de cuidados intensivos de um hospital português.

Verificou-se que os resultados obtidos para o modelo aditivo generalizado logístico não eram muito diferentes dos resultados obtidos pelo modelo aditivo generalizado com função de ligação baseada num perceptrão multicamada.Este facto pode estar relacionado com a dimensão da amostra utilizada, bem como pelo facto de as formas funcionais entre as variáveis explicativas e a variável resposta apresentarem efeitos quase lineares, o que faz com que não se consiga realçar a importância de uma função de ligação flexível em relação à logística. De acordo com os resultados obtidos, conclui-se que as melhorias de um modelo para o outro não são estatisticamente significativas, o que leva a concluir que não se justifica utilizar o modelo aditivo generalizado com função de ligação baseada num perceptrão multicamada, porque é menos parcimonioso.

O estudo comparativo efectuado neste trabalho deveria ser possível de realizar para um outro conjunto de dados, ou até um estudo de simulação, por forma a avaliar em que situações existem vantagens na utilização do modelo GAM-MLP.

Bibliografia

[1] Basak, T., Aciksoz, S., Tosun, B., Akyuz, A., & Acikel, C. (2013). Comparison of three different thermometers in evaluating the body temperature of healthy young adult individuals. *International journal of nursing practice*, 19 (5), 471-478.

[2] Bishop, C. M. (1996). Theorical foundations of neural networks. In *Physics Computing'96* (pp. 500-507). Academic Computer Centre.

[3] Bulhosa, M. C. (2019). *The impact of heatwaves on mortality in the Lisbon district – ICARO system revisited.* (Doctoral Dissertation). Faculdade de Ciências da Universidade de Lisboa.

[4] Cabral, M. S. (2019a). *Modelo Linear Generalizado Regressão Logística*, Slides. Faculdade de Ciências da Universidade de Lisboa.

[5] Cabral, M. S. (2019b). *Linear Models*, Slides. Faculdade de Ciências da Universidade de Lisboa.

[6] Cabral, M. S. (2019c). *Logistic Regressions Models*, Slides. Faculdade de Ciências da Universidade de Lisboa.

[7] Cabral, M. S. (2019d). *Modelo Linear Generalizado*, Slides. Faculdade de Ciências da Universidade de Lisboa.

[8] Cadarso-Suárez, C., Roca-Pardiñas, J., Figueiras, A., & González-Manteiga,W. (2005). Non-parametric estimation of the odds ratios for continuous exposures using generalized additive models with an unknown link function.*Statistics in medicine*, 24(8):1169–1183.

[9] Chaphalkar, N. B., Iyer, K. C.,& Patil, S. K. (2015). *Prediction of outcome of construction dispute claims using multilayer perceptron neural network model.* International Journal of Project Management, 33(8), 1827-1835.

[10] Cortes, C., & Mohri, M. (2003).AUC Optimization vs. Error Rate Minimization. *Advances in neural information processing systems*, 16, 313-320.

[11] Demler, O. V., Pencina, M. J., D'Agostino, R. B., Sr (2012). Misuse of DeLong test to compare AUCs for nested models. *Statistics in medicine*, 31(23), 2577–2587. https://doi.org/10.1002/sim.5328

[12] Faguet, G. B., & Davis, H. C. (1984). Regression Analysis in Medical Research. *Southern Medical Journal*, 77 (6), 722-5.

[13] Geraldes, C. (2017). Aplicação das Redes Neuronais Aditivas Generalizadas à Medicina. (Tese de Doutoramento). Faculdade de Ciências Médicas da Universidade Nova de Lisboa.

[14] Guisan, A., Edwards Jr, T. C., & Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributios: setting the scene. *Ecological Modelling*, 157 (2-3), 89-100.

[15] Hanusz, Z., Tarasinska, J., & Zielinski, W. (2016). Shapiro-Wilk test with known mean. *REVSTAT-Statistical Journal*, 14(1), 89-100.

[16] Hastie, T. J. (2017). *Generalized Additive Models* (pp. 249-307). Routledge.

[17] Hornik, K., Stinchcombe, M., & White, H. (1989).Multilayer feedforward networks are universal approximators. *Neural Networks*, 2 (5), 359-366.

[18] Lawrence, S., Giles, C. L., & Tsoi, A. C. (1998). What size neural network gives optimal generalization? Convergence Properties of Backpropagation. *Technical Reports*.

[19] Mackowiak, P. A., Wasserman, S. S., & Levine, M. M. (1992) . A critical appraisal of 98.6 F, the upper limit of the normal body temperature, and other legacies of Carl Reinhold August Wunderlich. *Jama*, 268 (12), 1578-1580.

[20] Mukaka M.(2012). Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal : the journal of Medical Association of Malawi*, 24(3), 69–71.

[21] Papoila, A. L., & Rocha, C. (2011). Modelling current status data using Generalized Additive Models with exible link: the additive gamma-logit model.*Int Journal of Applied MAth Statistics*, 24 (SI-11A), 2-19.

[22] Rencher, A. C., & Schaalje, G. B. (2008). *Linear Models in statistics*. John Wiley & Sons.

[23] Turkman, M. A. A., & Silva, G. L. (2000, September). Modelos Lineares Generalizados - da teoria à prática. In *VIII Congresso Anual da Sociedade Portuguesa de Estatística*, Edições SPE, Lisboa.

[24] Zhang, Z. (2018). Artificial neural network. *In Multivariate time series analysis in climate and environmental research* (pp. 1-35).

Apêndice

Script R utilizado para a realização do trabalho:

```
library(MASS)
library(gmodels)
library(aod)
library(lmtest)
library(Epi)
library(mgcv)
library(rms)
library(car)
library(caret)
library(rlang)
library(gam)
install.packages("mcgv")
library(mcgv)
install.packages("gam")
library(readxl)
library(sm)
install.packages("sm")
library(rms)
library(ResourceSelection)
library(car)
library(rmarkdown)
library(ggplot2)
library(ggpubr)
library(readxl)
library(nlme)
library(oddsratio)
library(pROC)
install.packages("boot",dep=TRUE)
library(boot)
install.packages("xtable")
```

```
library(xtable)
install.packages('scatterplot3d')
library(scatterplot3d)
install.packages('matrixStats')
library(matrixStats)
set.seed(123)
install.packages('matrixStats')
library(matrixStats)
set.seed(123)
or.lista<-NULL
attach(dados)


dados.originais<-dados
runs=seq(1:50)
for(run in runs){
  print(run)
  if(run==1){
    y<-dados[,5]
    y<-unlist(dados[,5])
    dados
    pts<-100
    x1<-dados[,1]
    min<-min(x1)
    max<-max(x1)
    media_x1<-mean(bodytemperature)
    dpadrao_x1<-sd(bodytemperature)
    x0 <-median(bodytemperature)
    inc<-(max-min)/pts
    dimensao<-nrow(dados)
  }else{
    indice<-sort(sample(1:dimensao,dimensao,replace=T),
    decreasing = F)
    dados<-dados.originais[indice,]
  }
  dados_0<-dados
  dados_0[,1]=x0

  formula = as.formula(death ~ s(bodytemperature) +
  s(bloodpressure ) + s(heartrate ) + s(serumpotassium ))

  gammlp<-function(dados,coloutcome,hn,formula,y){
```

```
  modelo.gam<- mgcv::gam(formula, family=binomial,
  data = dados)
  yint = as.vector(predict(modelo.gam,
  dados[,-coloutcome], type="response",
  se.fit = FALSE))
  print(yint)
  dados.treino<-data.frame(y,yint)
  modelo.mlp = neuralnet(y~yint, data=dados.treino,hidden=c(hn),
  linear.output=F)
  yhat = as.vector(modelo.mlp$net.result[[1]])
  list("yhat"=yhat,
  "modelo.mlp"=modelo.mlp,"yint"=yint,
  "modelo.gam"=modelo.gam,"dados.treino"=dados.treino)
}

gammlp.predict<-function(dados.new,modelo.gam,modelo.mlp){
  yint = as.numeric(predict(modelo.gam, dados.new,
  type="response", se.fit = FALSE))
  y<-rep(1,length(yint))
  aux<-data.frame(y,yint)
  yhat<-predict(modelo.mlp, aux)
  list("yint"=yint, "yhat"=yhat)
}


baseline<-gammlp(dados,coloutcome,34,formula,y)
  pi_0=gammlp.predict(dados_0[,-5],baseline$modelo.gam,
  baseline$modelo.mlp)$yhat
View(pi_0)

x<-min
z<-max
or<-NULL
x.plot<-NULL
dados_1<-dados
n<-dim(dados_1)[1]

repeat{
  dados_1[,1] = x
  pi_1 <- gammlp.predict(dados_1[,-5],baseline$modelo.gam,
  baseline$modelo.mlp)$yhat
  or_x = (1/n)*sum(((pi_1/(1-pi_1))/(pi_0/(1-pi_0))))
```

```
    print(or_x)
    x=x+inc
    or=c(or,or_x)
    x.plot<-c(x.plot,x)
    if(x>z) break;
  }
  print(pi_1)
  or.lista<-rbind(or.lista,or)
  print(or)
}

View(or.lista)
plot(x.plot,or)
print(or.lista)
q025f<-function(x,p) quantile(x,probs=0.25,na.rm = T)
medianaf<-function(x,p) quantile(x,probs=0.5,na.rm = T)
q075f<-function(x,p) quantile(x,probs=0.75,na.rm = T)

q025<-apply(or.lista,2,q025f)
mediana<-apply(or.lista,2,medianaf)
q075<-apply(or.lista,2,q075f)

plot(x.plot,q025,xlim=c(30,45),ylim=c(0,2),ylab = '',
xlab = '',col='red',lwd=2,lty=3,type='l')
abline(h=1,lwd=0.5)
abline(v=x0,lwd=0.5)
par(new=T)
plot(x.plot,mediana, xlim=c(30,45),ylim=c(0,2),ylab = '',
xlab = '',col='blue',lwd=2,type='l')
par(new=T)
plot(x.plot,q075, xlim=c(30,45),ylim=c(0,2),
xlab = 'Temperatura corporal',
ylab = '(ÔR)',col='red',lwd=2,lty=3,type='l')
par(new=F)
```